# Objectif

Le travail du statisticien consiste souvent à calculer des statistiques descriptives comparatives à partir de tris à plat ou tris croisés. L'intérêt de ces approches assez frustres est que la lecture de ces résultats ne nécessite pas de compétences particulières, elles permettent de comparer et de caractériser les spécificités d'une sous-population à partir d'une série de descripteurs.

Dans ce didacticiel, nous montrons l'utilisation de deux composants qui permettent de mettre en œuvre facilement ces opérations de comparaison entre sous-populations.

# Fichier

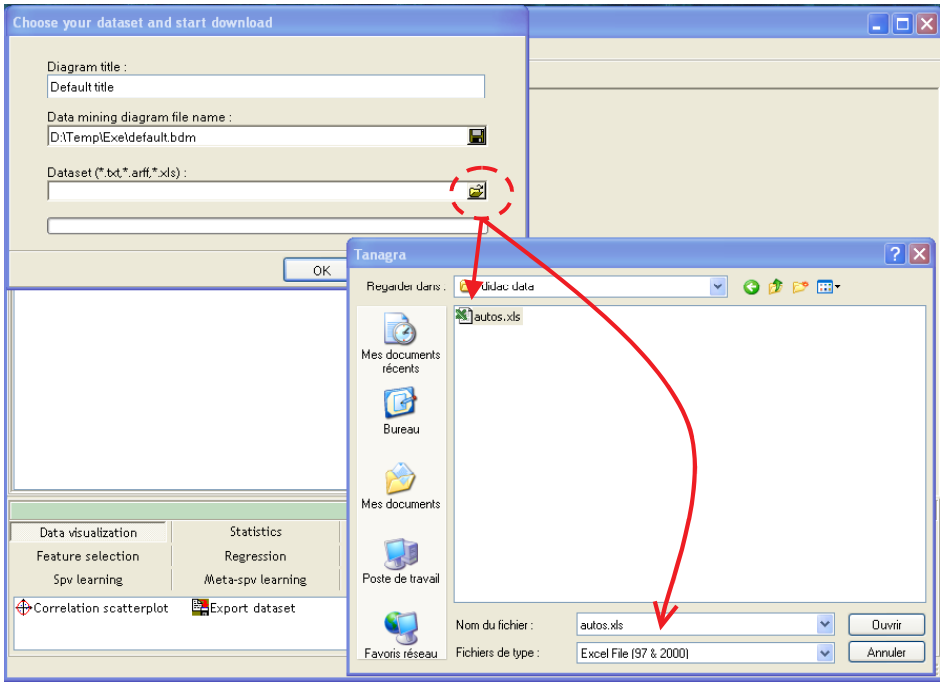
Nous travaillons sur le fichier « AUTOS.XLS » qui décrit 205 véhicules.

Nous voulons caractériser les véhicules selon leur consommation, leur prix, leur puissance et leur carrosserie en les regroupant selon leur type de carburant (FUEL-TYPE : GAS ou DIESEL) et le mode d'aspiration (ASPIRATION : STD ou TURBO).

# Construire les groupes et les caractériser

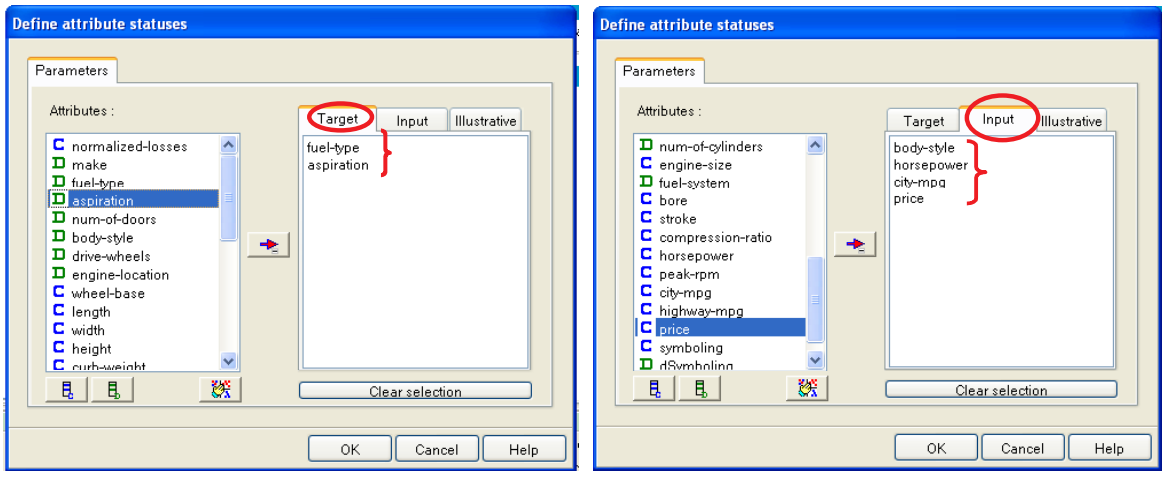
## Importer les données

Première étape toujours, importer le fichier de données. Pour cela, nous activons le menu FILE/NEW, et nous spécifions le format de fichier EXCEL.



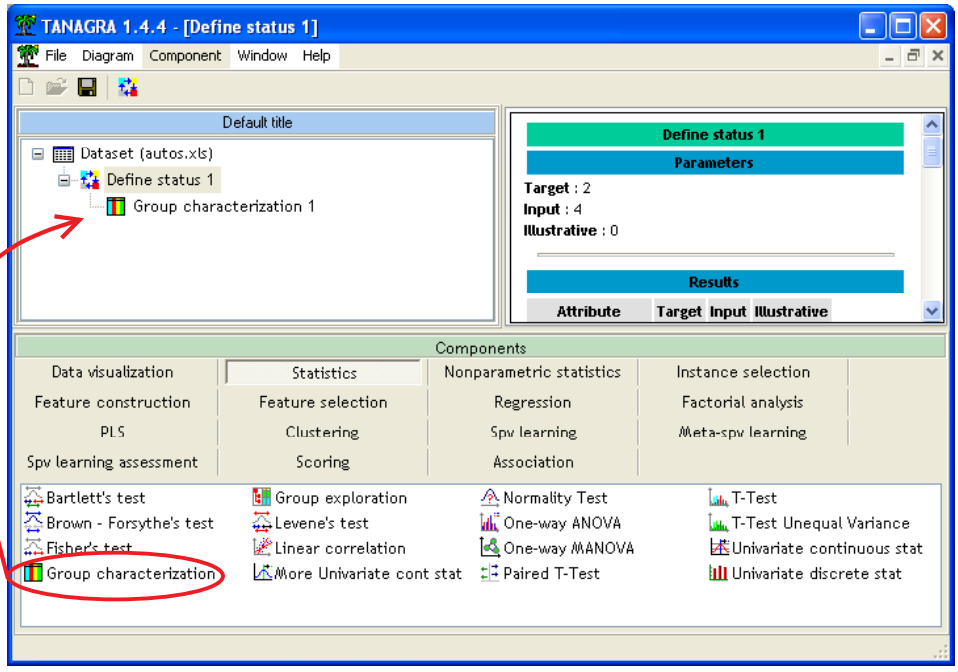
## Définir les attributs

Nous devons ensuite spécifier que nous voulons regrouper les individus selon le FUEL-TYPE et l'ASPIRATION : nous les plaçons donc en TARGET. Les descripteurs utilisés pour décrire les sous-groupes sont placés en INPUT, il s'agit de BODY-STYLE, HORSEPOWER, CITY-MPG et PRICE.




## TRIS A PLAT

Dans un premier temps, nous utilisons le composant GROUP CARACTERISATION qui utilise de manière indépendante chaque variable TARGET pour définir les sous-populations.



Ce composant produit des séries de tableaux, un pour chaque variable placée en TARGET. Les groupes sont donc formés de manière indépendante. Nous lisons dans le rapport la caractérisation des véhicules selon FUEL-TYPE, puis selon ASPIRATION.

**Description of "fuel-type"**

| fuel-type=gas | | | |
|---|---|---|---|
| Examples | | | [ 90.2 %] 185 |
| Att - Desc | Test value | Group | Overral |
| Continuous attributes | | | |
| horsepower | 2.4 | 106.40 | 104.26 |
| price | -1.6 | 12922.69 | 13207.13 |
| city-mpg | -3.6 | 24.67 | 25.22 |
| Discrete attributes | | | |
| body-style=hatchback | 2.9 | [ 98.6 %] 37.3 % | 34.1 % |
| body-style=convertible | 0.8 | [ 100.0 %] 3.2 % | 2.9 % |
| body-style=hardtop | -0.3 | [ 87.5 %] 3.8 % | 3.9 % |
| body-style=wagon | -0.4 | [ 88.0 %] 11.9 % | 12.2 % |
| body-style=sedan | -2.7 | [ 84.4 %] 43.8 % | 46.8 % |

| fuel-type=diesel | | | |
|---|---|---|---|
| Examples | | | [ 9.8 %] 20 |
| Att - Desc | Test value | Group | Overral |
| Continuous attributes | | | |
| city-mpg | 3.6 | 30.30 | 25.22 |
| price | 1.6 | 15838.15 | 13207.13 |
| horsepower | -2.4 | 84.45 | 104.26 |
| Discrete attributes | | | |
| body-style=sedan | 2.7 | [ 15.6 %] 75.0 % | 46.8 % |
| body-style=wagon | 0.4 | [ 12.0 %] 15.0 % | 12.2 % |
| body-style=hardtop | 0.3 | [ 12.5 %] 5.0 % | 3.9 % |
| body-style=convertible | -0.8 | [ 0.0 %] 0.0 % | 2.9 % |
| body-style=hatchback | -2.9 | [ 1.4 %] 5.0 % | 34.1 % |

**Description of "aspiration"**

| aspiration=std | | | |
|---|---|---|---|
| Examples | | | [ 82.0 %] 168 |
| Att - Desc | Test value | Group | Overral |
| Continuous attributes | | | |
| city-mpg | 2.9 | 25.84 | 25.22 |
| price | -2.5 | 12554.06 | 13207.13 |
| horsepower | -3.4 | 99.81 | 104.26 |
| Discrete attributes | | | |
| body-style=convertible | 1.2 | [ 100.0 %] 3.6 % | 2.9 % |
| body-style=hardtop | 0.4 | [ 87.5 %] 4.2 % | 3.9 % |
| body-style=hatchback | -0.1 | [ 81.4 %] 33.9 % | 34.1 % |
| body-style=sedan | -0.2 | [ 81.3 %] 46.4 % | 46.8 % |
| body-style=wagon | -0.3 | [ 80.0 %] 11.9 % | 12.2 % |

| aspiration=turbo | | | |
|---|---|---|---|
| Examples | | | [ 18.0 %] 37 |
| Att - Desc | Test value | Group | Overral |
| Continuous attributes | | | |
| horsepower | 3.4 | 124.43 | 104.26 |
| price | 2.5 | 16172.44 | 13207.13 |
| city-mpg | -2.9 | 22.41 | 25.22 |
| Discrete attributes | | | |
| body-style=wagon | 0.3 | [ 20.0 %] 13.5 % | 12.2 % |
| body-style=sedan | 0.2 | [ 18.8 %] 48.6 % | 46.8 % |
| body-style=hatchback | 0.1 | [ 18.6 %] 35.1 % | 34.1 % |
| body-style=hardtop | -0.4 | [ 12.5 %] 2.7 % | 3.9 % |
| body-style=convertible | -1.2 | [ 0.0 %] 0.0 % | 2.9 % |

**FUEL-TYPE** Nous constatons que 90.2% des véhicules sont à essence (GAS), leur puissance est légèrement au-dessus de la moyenne (106.4 hp contre 104.2 hp pour tous les véhicules), elles consomment plus (24.67 mpg[1] contre 25.22 mpg pour l'ensemble des véhicules).

[1] MPG est la norme anglo-saxonne, il s'agit du nombre de miles que l'on peut parcourir avec un gallon de carburant. Plus le chiffre est faible donc, plus la voiture consomme.

La colonne TEST VALUE[2] permet de quantifier l'intensité de la différence, il s'agit de la valeur normalisée d'un test de comparaison de moyenne. Il paraît illusoire de fixer un seuil qui permettrait de décider si la valeur test indique une différence statistiquement significative, il faut plutôt voir cet indicateur comme un critère qui permet de classer les variables selon l'intensité et le signe de l'écart avec la moyenne globale.

Concernant les véhicules DIESEL, nous constatons a contrario qu'ils consomment moins (30.30 mpg) et sont moins puissants (84.45 hp). Particularité par rapport aux véhicules essence, nous remarquons que les berlines (Carrosserie SEDAN) sont sur-représentés : elles sont 46.8% dans la totalité du fichier, elles représentent 75% des véhicules dans ce groupe [ $P(SEDAN / DIESEL) = 0.75$ , nous pouvons également interpréter cet indicateur comme une **précision**]. Nous pouvons résumer ces proportions dans le tableau croisé suivant.

| NB fuel-type | fuel-type | | |
|---|---|---|---|
| body-style | diesel | gas | Total |
| convertible | 0.00% | 3.24% | 2.93% |
| hardtop | 5.00% | 3.78% | 3.90% |
| hatchback | 5.00% | 37.30% | 34.15% |
| sedan | 75.00% | 43.78% | 46.83% |
| wagon | 15.00% | 11.89% | 12.20% |
| Total | 100.00% | 100.00% | 100.00% |

Autre lecture, nous constatons que le groupe des DIESEL représente 9.8% des véhicules, 15.6% des véhicules SEDAN se retrouvent dans ce groupe [ $P(DIESEL / SEDAN) = 0.156$ , nous pouvons interpréter cet indicateur comme un **rappel**]. La lecture du tableau croisé est alors inversée.

| NB fuel-type | fuel-type | | |
|---|---|---|---|
| body-style | diesel | gas | Total |
| convertible | 0.00% | 100.00% | 100.00% |
| hardtop | 12.50% | 87.50% | 100.00% |
| hatchback | 1.43% | 98.57% | 100.00% |
| sedan | 15.63% | 84.38% | 100.00% |
| wagon | 12.00% | 88.00% | 100.00% |
| Total | 9.76% | 90.24% | 100.00% |

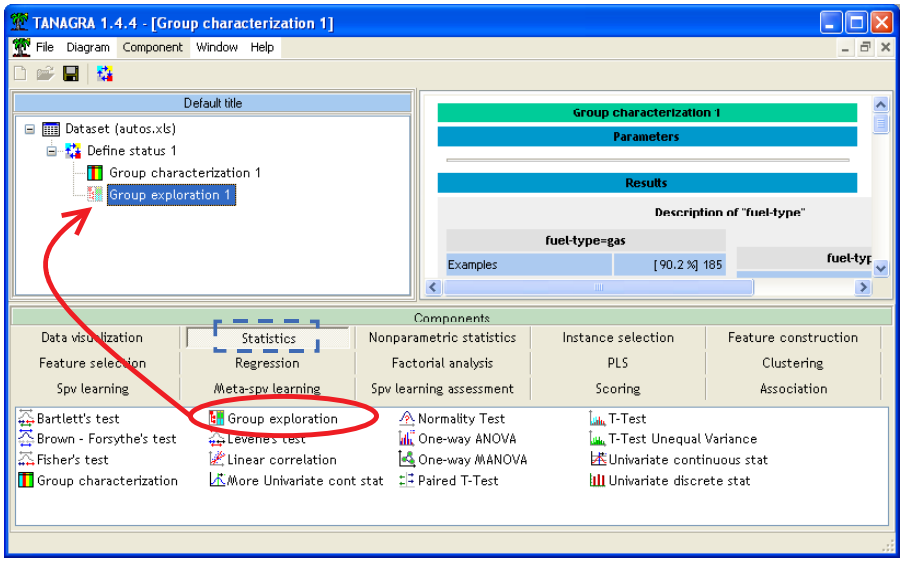
**ASPIRATION** Les véhicules TURBO représentent 18% de l'échantillon. Elles sont comparativement plus puissantes (124.43 hp), plus chères (16172 $) et consomment (22.41 mpg). Aucun type de carrosserie ne semble se démarquer dans ce groupe.

[2] LEBART, MORINEAU, PIRON, « Statistique exploratoire multidimensionnelle », DUNOD, 2000.

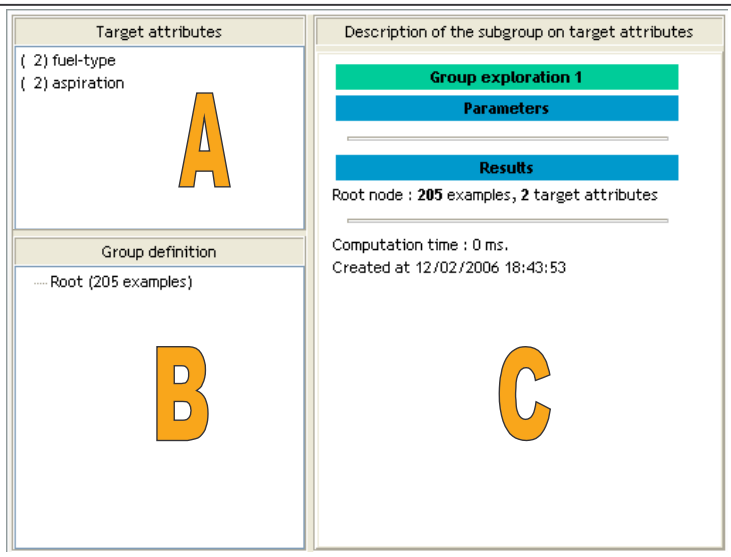
Nous pourrons mener une démarche analogue pour les véhicules STD. En revanche, et c'est la principale faiblesse de cet outil, nous ne pourrons pas effectuer d'analyse comparative en construisant des sous-groupes à partir des variables TARGET. Par exemple, il paraît difficile, à partir des résultats issus des tris à plat, de tirer des conclusions concernant les véhicules TURBO-DIESEL. Le composant **GROUP EXPLORATION** se propose justement de produire des **tris croisés sur plusieurs niveaux**, il peut mettre à contribution plusieurs variables TARGET. Il est possible de se focaliser sur un sous-groupe particulier en ne croisant que certaines modalités.

## TRIS CROISES SUR PLUSIEURS NIVEAUX

Nous voulons étudier les caractéristiques des véhicules TURBO-DIESEL. Pour ce faire, nous rajoutons le composant GROUP EXPLORATION dans le diagramme de traitements.

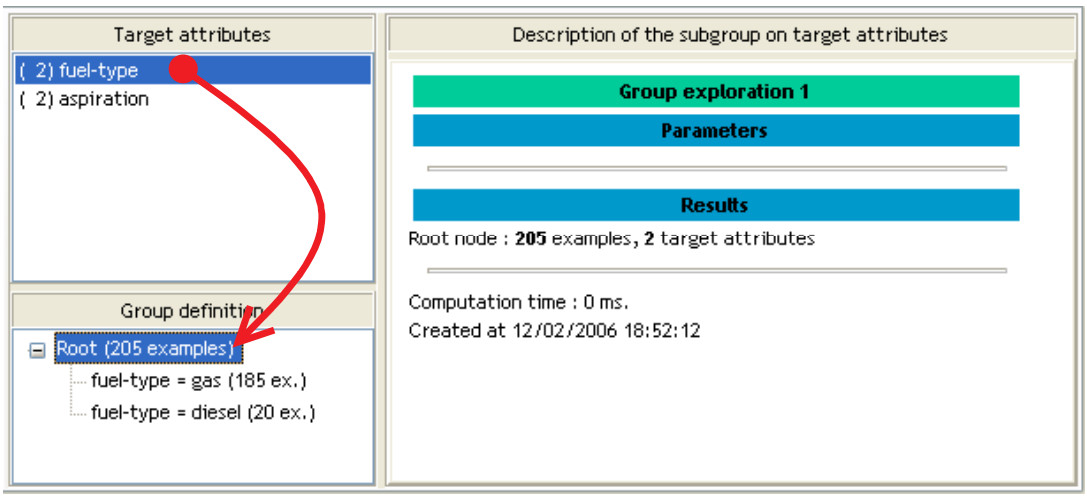



La fenêtre de visualisation est subdivisée en trois parties.

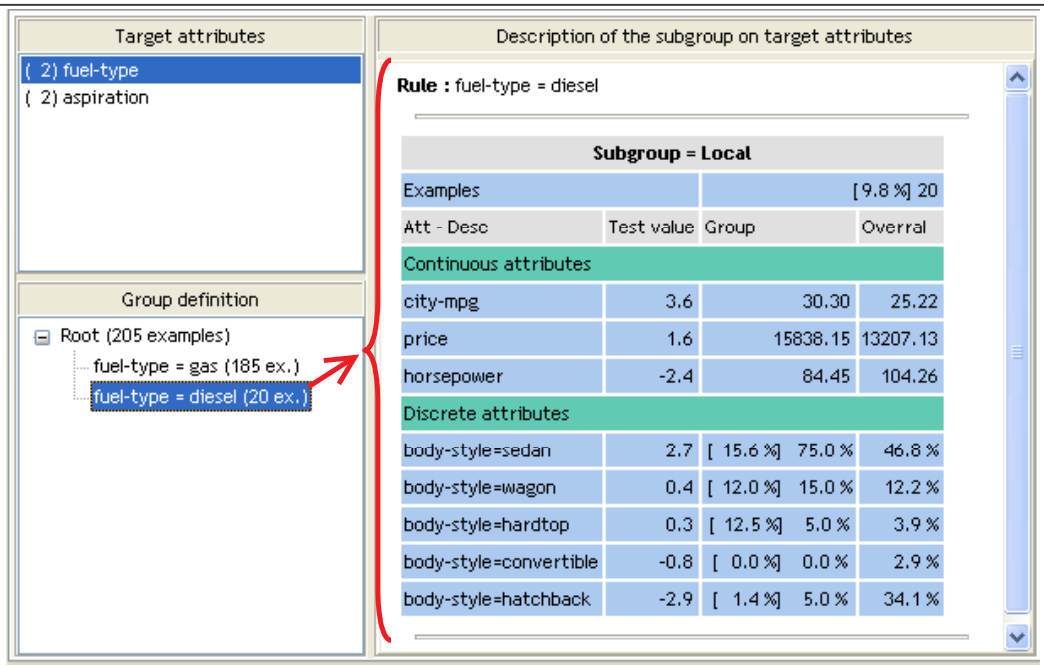



La partie [A – TARGET ATTRIBUTES] liste les variables qui servent à former les sous-groupes ; la partie [B – GROUP DEFINITION] représente sous une forme arborescente les sous-groupes qui sont formés par l'utilisateur ; enfin, la fenêtre [C – SUBGROUP DESCRIPTION] présente les statistiques relatives au groupe sélectionné dans la vue arborescente B.

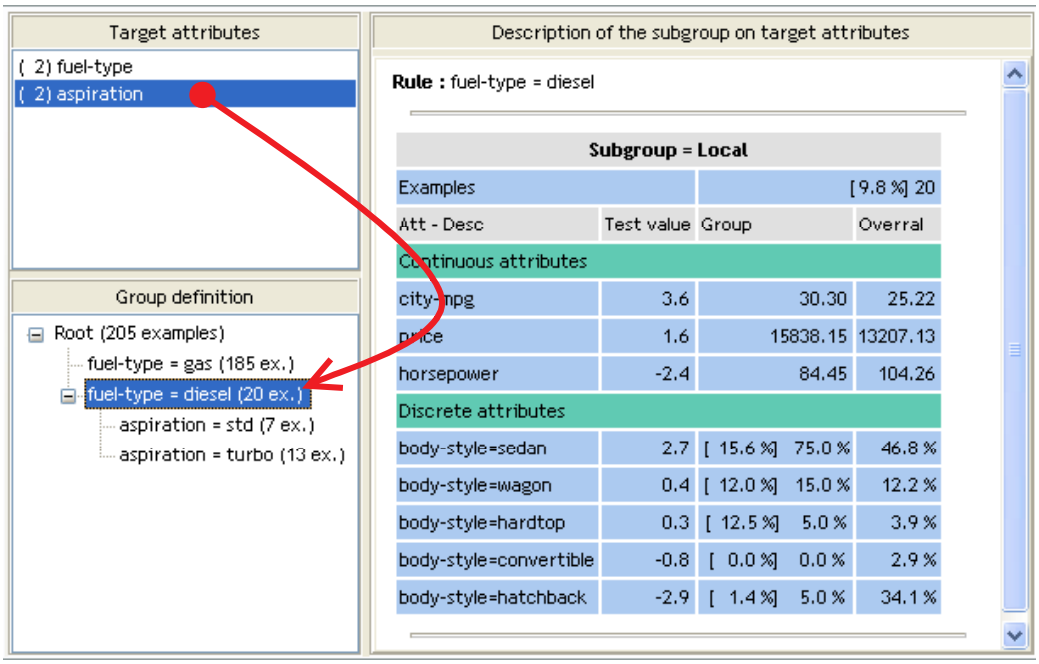
Nous constatons dans la racine de la vue arborescente que notre échantillon comporte 205 véhicules. Pour former les sous-groupes selon le type de carburant utilisé, il faut sélectionner la variable FUEL-TYPE, **la glisser et la déposer** sur le groupe que nous voulons explorer. Dans notre cas, il s'agit donc de le déposer sur la racine ROOT.



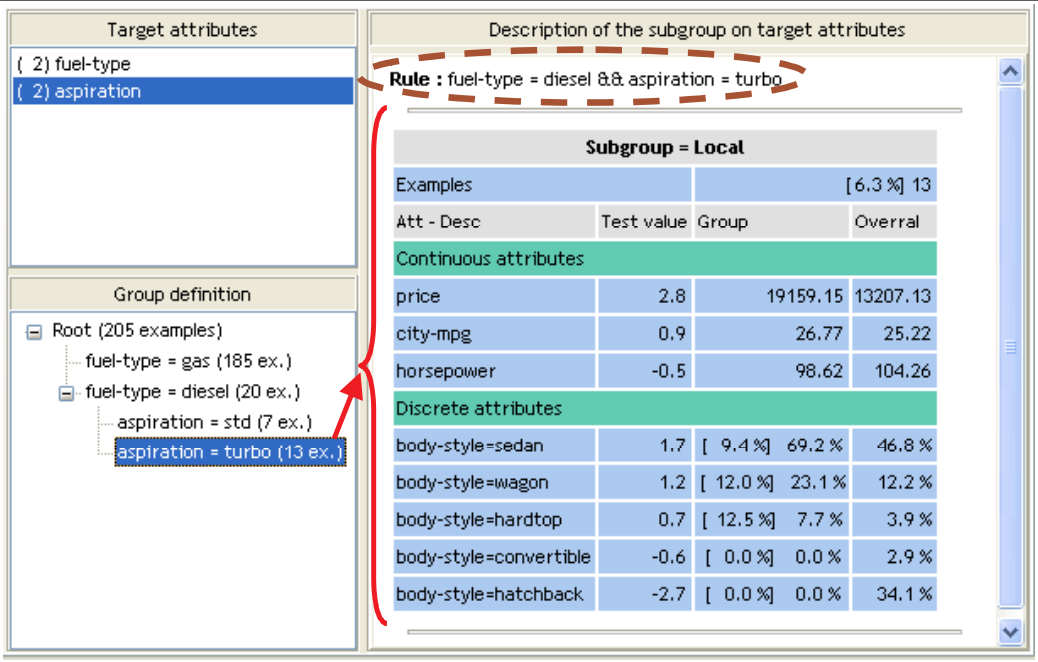

Nous retrouvons bien les proportions du premier composant, 90.2% (185) véhicules utilisent de l'essence (GAS), 9.8% (20) roulent au DIESEL. En sélectionnant le groupe des DIESEL, nous retrouvons également les mêmes statistiques comparatives.




Avec ce nouveau composant, il est possible maintenant d'explorer ce sous-groupe des véhicules diesels en la croisant avec les modalités de la variable ASPIRATION. Pour ce faire, toujours par glisser-déposer, nous insérons la variable ASPIRATION dans la fenêtre de définition des groupes.




Nous sélectionnons le groupe qui nous intéresse pour en visualiser les caractéristiques.




Il existe 13 (6.3%) véhicules TURBO-DIESEL dans la base. Nous constatons qu'ils sont plus chers que la moyenne globale des véhicules (19159 $), en revanche, ils ne se démarquent pas en termes de consommation (CITY-MPG) et de puissance (HORSEPOWER). A l'instar de ce que nous avions constaté pour les véhicules DIESEL, nous observons une sur-représentation du type de carrosserie SEDAN dans ce groupe.

**Bien entendu, il est possible de mettre beaucoup plus de variables dans les critères de tri (variables TARGET), le souci par la suite est la lisibilité des résultats, il est d'ailleurs possible de supprimer les sous-groupes qui ne présentent pas d'intérêt dans la vue arborescente. Petite restriction de ce composant, les variables TARGET doivent être exclusivement catégorielles.**